Graphical Models: Worksheet 2 MT 2023

Questions will not be marked, but solutions will be provided.

A: Warm Up

A1. Exponential Families. For each of the following, show that the family of distributions is an exponential family, and find the: (i) canonical and mean parameters; (ii) sufficient statistics; (iii) maximum likelihood estimate; (iv) cumulant function.

(a) The set of Binomial(n, p) distributions, with n fixed.
We have $\log L(p; x) = l(p; x) = x \log p + (n - x) \log(1 - p)$ *(ignoring constants) giving*

$$l(p; x) = x \log \frac{p}{1-p} + n \log(1 - p).$$

Hence the sufficient statistic is $\phi(x) = x$ *and canonical parameter is* $\theta = \log \frac{p}{1-p} = \operatorname{logit} p$. *To find the cumulant function, note that*

$$p = \frac{e^\theta}{1 + e^\theta}$$
$$\log(1 - p) = -\log(1 + e^\theta)$$

so $A(\theta) = \log(1 + e^\theta)$. *Then* $A'(\theta) = n\frac{e^\theta}{1+e^\theta} = np$ *is the mean parameter (and equals* $\mathbb{E}X$ *as expected), and the MLE is* $\hat{p} = x/n$. *Finally,* $A''(\theta) = n\frac{e^\theta(1+e^\theta)-e^{2\theta}}{(1+e^\theta)^2} = n\frac{e^\theta}{(1+e^\theta)^2} = np(1 - p)$ *which is* $\operatorname{Var} X$ *as expected.*

(b) The set of Gamma distributions with parameters (a, b).
This time

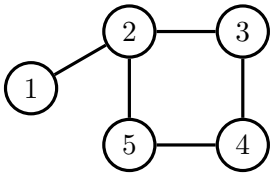
$$\begin{aligned} l(a, b; x) &= (a - 1) \log x - bx + a \log b - \log \Gamma(a) \\ &= a \log x - bx + a \log b - \log \Gamma(a) - \log x \end{aligned}$$

which suggests canonical parameters a, b *and sufficient statistics* $\phi_1(x) = \log x$ *and* $\phi_2(x) = -x$. *We have* $A(\theta) = -a \log b + \log \Gamma(a)$ *directly, and* $A'(\theta) = (\Gamma'(a)/\Gamma(a) - \log b,\ a/b)$. *One can easily check that* $\mathbb{E}X = a/b$, *and indeed it is also the case that* $\mathbb{E} \log X = \Gamma'(a)/\Gamma(a) - \log b$.

For (a), show directly that the derivatives of the cumulant function give the first two centred moments of the sufficient statistics (see Lemma 3.1).

A2. Graphical Separation. Consider the graph below. List all the independences implied by the pairwise Markov property.

Give one conditional independence that follows from the global Markov property but is not already in your list.



The pairwise property gives

$$X_1 \perp\!\!\!\perp X_3 \mid X_2, X_4, X_5 \qquad X_1 \perp\!\!\!\perp X_4 \mid X_2, X_3, X_5$$
$$X_2 \perp\!\!\!\perp X_4 \mid X_1, X_3, X_5 \qquad X_3 \perp\!\!\!\perp X_5 \mid X_1, X_2, X_4$$
$$X_1 \perp\!\!\!\perp X_5 \mid X_2, X_3, X_4$$

The global implies, for example, that $X_1 \perp\!\!\!\perp X_3, X_4 \mid X_2, X_5$, which can only be obtained from the pairwise property by using property 5 of the graphoid axioms.

B: Core Questions

B1. Markov Properties.

Let $\mathcal{G}$ be a graph, and define the *boundary* of a vertex v by

$$\mathrm{bd}_{\mathcal{G}}(v) \equiv \{w \in V \setminus \{v\} \mid w \sim v\}.$$

A distribution obeys the *local Markov property* with respect to $\mathcal{G}$ if

$$X_v \perp\!\!\!\perp X_{V \setminus (\mathrm{bd}_{\mathcal{G}}(v) \cup \{v\})} \mid X_{\mathrm{bd}_{\mathcal{G}}(v)}, \qquad \forall v \in V.$$

(a) Show that if p obeys the local Markov property then this implies that p obeys the pairwise Markov property.
Suppose that $v \not\sim w$; then $w \in V \setminus (\mathrm{bd}_{\mathcal{G}}(v) \cup \{v\})$. Hence applying the local Markov property gives $X_v \perp\!\!\!\perp X_w, X_{V \setminus (\mathrm{bd}_{\mathcal{G}}(v) \cup \{v,w\})} \mid X_{\mathrm{bd}_{\mathcal{G}}(v)}$, and note that using property 3 from the graphoid axioms gives

$$X_v \perp\!\!\!\perp X_w \mid X_{\mathrm{bd}_{\mathcal{G}}(v)} \cup X_{V \setminus (\mathrm{bd}_{\mathcal{G}}(v) \cup \{v,w\})}$$
$$X_v \perp\!\!\!\perp X_w \mid X_{V \setminus \{v,w\}}.$$

(b) Show that the global Markov property implies the local Markov property.
Clearly any path from v to $w \in V \setminus (\mathrm{bd}(v) \cup \{v\})$ must pass through $\mathrm{bd}(v)$ at its vertex adjacent to v. Hence $\mathrm{bd}_{\mathcal{G}}(v)$ separates v from $V \setminus (\mathrm{bd}_{\mathcal{G}}(v) \cup \{v\})$, so the global Markov property gives us

$$X_v \perp\!\!\!\perp X_{V \setminus (\mathrm{bd}_{\mathcal{G}}(v) \cup \{v\})} \mid X_{\mathrm{bd}_{\mathcal{G}}(v)}$$

as required.
The local gives

$$X_1 \perp\!\!\!\perp X_3, X_4 \mid X_2, X_5 \qquad X_2 \perp\!\!\!\perp X_4 \mid X_1, X_3, X_5$$
$$X_4 \perp\!\!\!\perp X_1, X_2 \mid X_3, X_5 \qquad X_3 \perp\!\!\!\perp X_1, X_5 \mid X_2, X_4$$
$$X_5 \perp\!\!\!\perp X_3 \mid X_1, X_2, X_4.$$

(c) Show that, if p is strictly positive and obeys the pairwise Markov property with respect to $\mathcal{G}$, then p also obeys the global Markov property with respect to $\mathcal{G}$.
[Hint: Property 5 of the graphoid axioms and the proof of Theorem 4.10 may be helpful.]
Suppose A is separated from B by S. Let $\tilde{A}$ be the set of vertices connected to A by paths in $\mathcal{G}_{V \setminus S}$, and $\tilde{B} = V \setminus (\tilde{A} \cup S)$; clearly $\tilde{A}$ is separated from $\tilde{B}$ by S, so there is no edge between a, b for any $a \in \tilde{A}$ and $b \in \tilde{B}$. Repeatedly applying property 5 and (using the fact that p is positive) gives $X_{\tilde{A}} \perp\!\!\!\perp X_{\tilde{B}} \mid X_S$. Then applying graphoid property 2 shows that $X_A \perp\!\!\!\perp X_B \mid X_S$.

(d) Give an example of a graph in which property 5 is required for the pairwise Markov property to imply the local Markov property. Hence or otherwise find a distribution in which the pairwise property holds with respect to this graph, but the local property does not.

The graph above does this, but a simpler example is just to have three vertices X, Y, Z with no edges, and all equal with probability 1. Then certainly $X \perp\!\!\!\perp Y \mid Z$ (and permutations thereof) but also clearly $X \not\perp\!\!\!\perp Y, Z$ as required by the local Markov property, provided that the distribution of X is not degenerate.

B2. Decomposability

Complete the proof of Theorem 4.20 from lectures; that is, show that if $\mathcal{G}$ is an undirected graph, (iii) the fact that every minimal a, b-separator is complete implies that (iv) the cliques satisfy the running intersection property starting with a given C; and that (iv) implies (i): the graph is decomposable.

(iii) $\implies$ (iv). We use induction on $m = |V|$; if the graph is complete there is nothing to prove, so the result holds for $m \leq 1$. Otherwise pick a, b not adjacent and let S be a minimal separator. As in Theorem 4.10, let $\tilde{A}$ be the connected component of a in $\mathcal{G}_{V \setminus S}$, and $\tilde{B} = V \setminus (S \cup \tilde{A})$ be the rest. Any minimal separators in the induced subgraphs $\mathcal{G}_{\tilde{A} \cup S}$ and $\mathcal{G}_{\tilde{B} \cup S}$ are no larger than in $\mathcal{G}$, so will also be complete, and hence applying the result by induction on m gives two sequences of cliques that satisfy running intersection. The set S is complete in both subgraphs, so there is some clique D in $\mathcal{G}_{\tilde{A} \cup S}$ that contains S. In addition, each clique in $\mathcal{G}$ is a clique in one of the two subgraphs. Hence it is easy to see that if we order the cliques to satisfy running intersection for $\mathcal{G}_{\tilde{A} \cup S}$ and $\mathcal{G}_{\tilde{B} \cup S}$ respectively (or the other way around depending on which graph C is contained in), and being sure to start with a clique of $\mathcal{G}_{\tilde{B} \cup S}$ that contains S, then together they will satisfy running intersection for $\mathcal{G}$, starting with C.

(iv) $\implies$ (i). As suggested in the notes, we proceed by induction on the number of cliques k; if $k = 1$ there is nothing to prove. Let $H_{k-1} = C_1 \cup \cdots \cup C_{k-1}$, and $S_k = C_k \cap H_{k-1}$, and $R_k = C_k \setminus S_k$; we must show that $(H_{k-1} \setminus S_k, S_k, R_k)$ is a proper decomposition.

Well certainly S_k is complete, and if R_k is empty then that would imply that C_k is contained in one of the C_i for $i < k$, so we could use the induction hypothesis for $k - 1$ cliques. Otherwise, we just need to show that there are no edges between $H_{k-1} \setminus S_k$ and R_k. Suppose for contradiction that $h - r$ is such an edge: this edge must be contained within some clique C_i; but $R_k = C_k \setminus \bigcup_{i<k} C_i$ (including c) are precisely the vertices not contained in any previous clique, so the only possible clique to contain this edge is C_k. However $H_{k-1} \setminus S_k = H_{k-1} \setminus C_k$ is disjoint from C_k, and hence there is no such edge.

Now we have a proper decomposition, and the graph $\mathcal{G}_{H_{k-1}}$ has $k-1$ cliques $C_1, \ldots, C_{k-1}$ that satisfy the running intersection property, so by the induction hypothesis this subgraph is decomposable; $\mathcal{G}_{C_k}$ is complete. Hence (by definition), the original graph $\mathcal{G}$ is decomposable.

Does decomposability imply that every minimal A, B-separator is complete, for sets A and B?

No, in a graph like $1-2-3-4-5$, the minimal separator of $\{1, 5\}$ and $\{3\}$ is $\{2, 4\}$, which isn't complete.

B3. Whittaker Data

Using `R`, load the Whittaker data from lectures with the commands:

```
> library(ggm)
> data(marks)
> head(marks, 8)  # inspect the first few
> solve(cov(marks))  # empirical concentration matrix
```

(if not already installed, you may have to call `install.packages("ggm")`, to use the `ggm` package).

(a) Manually find the MLE for the covariance matrix Σ, under the model from lectures in which 'analysis' and 'statistics' are independent of 'mechanics' and 'vectors' conditional on 'algebra'.

[Hint: R commands you might need are `solve()`, which inverts matrices, and the use of square brackets `[]` for subsetting. See the MSc R Programming lecture notes for details.]

(b) Suppose we have i.i.d. observations $x_V^{(1)}, \ldots, x_V^{(n)}$ from a multivariate Gaussian with known mean μ and unknown covariance Σ. Show that the log-likelihood for Σ can be written as

$$l(\Sigma; X) = -\frac{n}{2}\log|\Sigma| - \frac{n}{2}\operatorname{tr}(S\Sigma^{-1}).$$

where $S = \frac{1}{n}\sum_{i=1}^{n}(x_V^{(i)} - \mu)(x_V^{(i)} - \mu)^T$ and tr() is the trace operator.
[Hint: $\operatorname{tr}(AB) = \operatorname{tr}(BA)$*]*

Up to a constant the log-likelihood is

$$l(\Sigma; X) = -\frac{n}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{n}(x_V^{(i)} - \mu)^T\Sigma^{-1}(x_V^{(i)} - \mu).$$

Now, since the term in the sum is a scalar it is the same as its own trace; using the hint gives

$$\begin{aligned} l(\Sigma; X) &= -\frac{n}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{n}\operatorname{tr}(x_V^{(i)} - \mu)(x_V^{(i)} - \mu)^T\Sigma^{-1} \\ &= -\frac{n}{2}\log|\Sigma| - \frac{n}{2}\operatorname{tr}\left\{\frac{1}{n}\sum_{i=1}^{n}(x_V^{(i)} - \mu)(x_V^{(i)} - \mu)^T\Sigma^{-1}\right\} \end{aligned}$$

as required.

(c) Hence carry out a likelihood ratio test to see whether this model is a good fit to the data.

The model has four zero correlations so we should compare twice the difference in the log-likelihoods with a χ^2_4 distribution.

This is not significantly large, since a χ^2_4 has mean 4 and variance 8; hence the model is a good fit. The p-value for the test is:

B4. Iterative Proportional Fitting

(a) Show that the iterative proportional fitting algorithm does not decrease the likelihood at any iteration. Argue also that, if the likelihood does not strictly increase after a full cycle of updates then the algorithm has converged to a solution.

Pick a margin A, and let $B = V \setminus A$. The algorithm replaces

$$p^{(t)}(x_A, x_B) = p^{(t)}(x_A) \cdot p^{(t)}(x_B \mid x_A) \mapsto q(x_A) \cdot p^{(t)}(x_B \mid x_A)$$

Then note that the log-likelihood is

$$\begin{aligned} l(p; n) &= \sum_{x_V} n(x_V) \log p^{(t)}(x_A, x_B) \\ &= \sum_{x_A} n(x_A) \log p^{(t)}(x_A) + \sum_{x_V} n(x_V) \log p^{(t)}(x_B | x_A) \end{aligned}$$

and notice that the first term is maximized by the IPF step, but the second term is unchanged.

Consider a 7-dimensional table, and suppose that we have an undirected model based on the cliques $\{1, 2, 3\}$, $\{1, 2, 4\}$, $\{2, 3, 5\}$, $\{1, 3, 6\}$ and $\{5, 7\}$.

(b) Show that this model is decomposable and that if the IPF algorithm is run in the order given above, it will return the MLE after a single iteration of updating each clique in turn.

The sets above satisfy the running intersection property in the order given, and therefore the model is decomposable.

The rest of the solution gives much more detail than is really required. We start with a uniform distribution, $p^{(0)}(x_V) = \prod_{v \in V} |\mathfrak{X}_v|^{-1}$. We claim that after $i - 1$ updates we will have:

$$p^{(i-1)}(x_V) = \prod_{j=1}^{i-1} p(x_{R_j} \mid x_{S_j}) \cdot \prod_{w \in V \setminus H_{i-1}} \frac{1}{|\mathfrak{X}_w|}, \tag{1}$$

where $R_i = C_i \setminus S_i$ and S_i is the separator set, and $H_{i-1} = \bigcup_{j<i} C_i$ is the history. *This will then be replaced by*

$$p^{(i)}(x_V) = p^{(i)}(x_V) \frac{p(x_{C_i})}{p^{(i-1)}(x_{C_i})}.$$

Notice that, by an application of Theorem 4.24 to the graph with the first $i - 1$ cliques only, the margin of $p^{(i-1)}$ over H_{i-1} is just $p(x_{H_{i-1}})$, and therefore its margin over S_i is also just $p(x_{S_i})$. On the other hand, nothing in R_i has been updated yet, so its distribution will still be uniform under $p^{(i-1)}$. Hence

$$p^{(i)}(x_V) = p^{(i-1)}(x_{H_{i-1} \setminus S_i} \mid x_{S_i}) \cdot p(x_{S_i}) \cdot \frac{p(x_{C_i})}{p(x_{S_i}) \cdot \prod_{w \in R_i} |\mathfrak{X}_w|^{-1}} \prod_{w \in V \setminus H_i} \frac{1}{|\mathfrak{X}_w|}.$$

This is equivalent to the ith version of (1), so by induction we have that indeed (after performing all the updates), we obtain:

$$\begin{aligned} p(x_V) &= \prod_{i=1}^{k} p(x_{R_i} \mid x_{S_i}) \\ &= p(x_{123}) \cdot p(x_4 \mid x_{12}) \cdot p(x_5 \mid x_{23}) \cdot p(x_6 \mid x_{13}) \cdot p(x_7 \mid x_5). \end{aligned}$$

It then follows from the comments below Theorem 4.24 that this is indeed the MLE.

(c) Is the same true if we choose the order $\{1,2,4\}$, $\{2,3,5\}$, $\{1,3,6\}$, $\{5,7\}$ and $\{1,2,3\}$?

No. This order of updates does not respect any junction tree (see Section 7), so for most distributions this will not converge in a single sequence of updates. You may like to verify this fact numerically.

C: Optional

C1. Marginal Models

Let $\mathcal{G}$ be a graph containing a path $\pi : i - k_1 - \cdots - k_m - j$, for $m \geq 0$.

(a) Construct a distribution p that factorizes according to $\mathcal{G}$, and such that for any set $C \subseteq V \setminus \{i, j, k_1, \ldots, k_m\}$ we have $X_i \not\perp\!\!\!\perp X_j \mid X_C\ [p]$.

[Hint: remember that undirected graphs generalize Markov chain models.]

A simple example would be to construct a Markov chain along the path. (Abusing notation slightly, let $X_t = X_{k_t}$, with $X_0 = X_i$ and $X_{m+1} = X_j$.) Let $X_0 \sim N(0,1)$ and let

$$X_{i+1} = \rho X_i + \sqrt{1-\rho^2} Z_i, \ldots i = 0, \ldots, m,$$

where $Z_i \sim N(0,1)$ independently, and $\rho \in (0,1)$. Then one can check that $\mathrm{Cov}(X_t, X_{t+s}) = \rho^{|s|} \neq 0$, for example.

Now, assume that all other variables are completely independent of the variables on the path, so that p factorizes according to the graph whose cliques are just the edges in π. Then, for example, the conditional covariance $\mathrm{Cov}(X_t, X_{t+s} \mid X_C) = \mathrm{Cov}(X_t, X_{t+s})$ is just the same as the marginal covariance. This gives the required result.

(b) Deduce that for any graph $\mathcal{G}$ and sets A, B, S such that $A \not\perp_s B \mid S$ in $\mathcal{G}$, there exists a distribution p which factorizes according to $\mathcal{G}$ and for which $X_A \not\perp\!\!\!\perp X_B \mid X_S$ in p. (In this sense the global Markov property is *complete*; separation represents **all** the independences guaranteed by factorization.)

Since the separation does not hold, pick any $a \in A$, $b \in B$ joined by a path π in $\mathcal{G}$ that does not pass through S. Now construct the distribution as above, and note that $X_a \not\perp\!\!\!\perp X_b \mid X_S$; hence $X_A \not\perp\!\!\!\perp X_B \mid X_S$.

In fact, using facts about polynomials one can show that for each graph there exists a single distribution p such that $A \perp_s B \mid S$ in $\mathcal{G}$ if and only if $X_A \perp\!\!\!\perp X_B \mid X_S$ under p; this is beyond the scope of the course.

Given an undirected graph $\mathcal{G}$ and subset of vertices $W \subseteq V$, define $\mathcal{G}|^W$ as the undirected graph with vertex set W, and an edge $i - j$ if and only if there is a path from i to j in $\mathcal{G}$ with all intermediate vertices in $V \setminus W$. [Note that this is quite different to the induced subgraph $\mathcal{G}_W$.]

(c) Let $p(x_V)$ be globally Markov with respect to $\mathcal{G}$. Show that $p(x_W) = \sum_{x_{V \setminus W}} p(x_V)$ is globally Markov with respect to $\mathcal{G}|^W$.

Consider a pair of vertices a, b, and a conditioning set $S \subseteq W \setminus \{a, b\}$. First note that if there is an open path given S between two vertices in $\mathcal{G}$ then there is also a corresponding path in $\mathcal{G}|^W$, and this path uses only a (potentially) subset

of the vertices of the original path, since it skips all those in $V \setminus W$. Hence, by the contrapositive of this statement, if there is a separation between a and b by S in $\mathcal{G}|^W$ (because there are no open paths), there must also be such a separation in $\mathcal{G}$. Then by the global Markov property applied to $\mathcal{G}$ we have $X_a \perp\!\!\!\perp X_b \mid X_S$ under $p(x_V)$. Clearly then this also applies to the marginal distribution $p(x_W)$.

(d) Show further that, in general, $p(x_W)$ is **not** globally Markov with respect to any edge subgraph of $\mathcal{G}|^W$.

Suppose we have a graph $\mathcal{H}$ over W such that $a \not\sim b$ in $\mathcal{H}$ but $a \sim b$ in $\mathcal{G}|^W$. This means that there is a path in $\mathcal{G}$ passing through only vertices in $V \setminus W$, and by (a) we can construct a distribution that makes X_a and X_b correlated but independent of all other variables. In particular, $X_a \not\perp\!\!\!\perp X_b \mid X_{W\setminus\{a,b\}}$ and p is not pairwise Markov with respect to $\mathcal{H}$.

C2. Hierarchical Models.

Let $\mathcal{C}$ be a collection of non-empty subsets of a set V, such that:

- $\bigcup_{C\in\mathcal{C}} C = V$;
- for any distinct $C, D \in \mathcal{C}$ we have $C \not\subset D$.

In other words, this is a set of inclusion maximal subsets. We call $\mathcal{C}$ a *generating class*.

(a) Show that the cliques of a graph are a generating class.

Each vertex $\{v\}$ is complete, so must be contained within some maximal complete set (i.e. clique). By definition, cliques are maximal, so the second condition is satisfied.

(b) List, up to symmetry, all the generating classes on the set $V = \{1, 2, 3\}$. Do all generating classes correspond to the cliques of a graph?

Apart from singletons, the only generators consist of those with 1, 2, and 3 subsets of size 2, as well as $\{1, 2, 3\}$ (for a total of 4). The simplest non-graphical example is $\{1, 2\}$, $\{2, 3\}$, $\{1, 3\}$, which can't correspond to a graph because every pair is contained in a subset but $\{1, 2, 3\}$ is not.

Up to isomorphism, this gives: $\{1\}$, $\{2\}$, $\{3\}$
$\{1, 2\}$, $\{3\}$
$\{1, 2\}$, $\{2, 3\}$
$\{1, 2\}$, $\{2, 3\}$, $\{1, 3\}$
$\{1, 2, 3\}$

Given a generating class $\mathcal{C}$, we can define a corresponding log-linear model by requiring that $\lambda_A = 0$ whenever A is not a subset of any element of $\mathcal{C}$. Such models are called *hierarchical*.

(c) Show that the counts $n(x_C)$ for $C \in \mathcal{C}$ are sufficient statistics for this model.

The data below consist of answers from high schoolers to a Dayton, Ohio survey on substance use.

Alcohol	Tobacco	Marijuana	
		Yes	No
Yes	Yes	911	538
	No	44	456
No	Yes	3	43
	No	2	279

They are available in a text file, `substance.txt`, on the class website. After downloading the file, you can read these data into R using

```
> dat <- read.table("substance.txt", header = TRUE)
```

(d) Using `glm()` with `family=poisson`, fit a hierarchical model to these data with the generating class $\mathcal{C} = \{\{A, M\}, \{T, M\}, \{A, T\}\}$. Is it a good fit? Is any smaller hierarchical model a good fit?

The residual deviance is small compared to the number of degrees of freedom, so we would say this model is a good fit to the data.

Trying any sub-model (which would also be a graphical model) gives a bad fit. The least bad omits the alcohol-marijuana term, and gives a deviance of 92 on 2 degrees of freedom.

(e) Verify that the fitted distribution has the same sufficient statistics as the data.

(f) Try adding the vector $(+1, -1, -1, +1, -1, +1, +1, -1)$ to your counts. Verify that the parameter estimates are unchanged with this 'new data'. Can you explain why? *The sufficient statistics are unchanged, hence so is the likelihood and the inference. Note, however, that the model* fit *has changed!*